BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 8: e138931
doi: 10.3897/biss.8.138931

OPEN ACCESS

Conference Abstract

SQLite: A “Frictionless” Solution for Exchange of Biodiversity Data?

Dmitry Mozzherin[‡], Geoffrey Donald Ower[‡]

[‡] University of Illinois, Champaign, Illinois, United States of America

Corresponding author: Dmitry Mozzherin (dmozzherin@gmail.com)

Received: 09 Oct 2024 | Published: 10 Oct 2024

Citation: Mozzherin D, Ower G (2024) SQLite: A “Frictionless” Solution for Exchange of Biodiversity Data? Biodiversity Information Science and Standards 8: e138931. https://doi.org/10.3897/biss.8.138931

Abstract

Biodiversity data exchange depends on established standards, such as Darwin Core, Audiovisual Core, Taxon Concept Schema, etc. Standards provide terms with defined semantic meaning and structures for data exchange. Standards simplify interchange of biodiversity data among government agencies, researchers, and engineers.

A notable challenge during such data transfers is the complexity involved in processing data for various uses. Most data exchanges are performed via Comma Separated Value (CSV), Extensible Markup Language (XML), or JavaScript Object Notation (JSON), and need to be parsed and imported into a database before being queryable. For example, to explore the data in a Darwin Core Archive zipped file, it is necessary to decompress it, parse XML-based metadata about the content of the file, parse and read Ecological Metadata Language (EML) to extract provenance metadata, and correctly open the text-delimited files that use a wide variety of character encodings, delimiters, enclosures, and escape characters. All of it requires non-trivial data management and programming skills from users. We propose a paradigm shift toward using queryable SQLite-based files for more straightforward data interchange. This approach would reduce friction in data processing by directly using Structured Query Language (SQL).

SQLite is an open source, high performance, lightweight database already installed on most computers (SQLite 2024). It integrates a database into a single file, facilitating straightforward data exchange and compression. Its robust engine is able to manage terabytes of data. The SQLite developers are committed to maintaining backward

compatibility of both binary and SQL textfile versions until 2050 (SQLite 2024). Connectivity to SQLite databases is supported by all popular programming languages. Furthermore, the United States Library of Congress endorses SQLite alongside XML and JSON for data archives, attesting to its long-term reliability (United States Library of Congress 2024).

We at the Species File Group are experimenting with using SQLite to create a universal data converter, in which an SQLite database serves as an intermediate data storage format. This provides several useful advantages:

1. Universal data converter

By learning a standard SQL schema for biodiversity data, users are able to efficiently write importers and exporters to/from the intermediate schema for a variety of other biodiversity data-exchange formats. Any programming language can be used to develop the conversion scripts. Import and export are completely decoupled from each other, allowing mixing and matching of converters in a flexible way. For example, we aim to use SQLite as an intermediary archive to convert data between Darwin Core Archives, Catalogue of Life Data Package (ColDP), Global Names Verifier, and TaxonWorks (Fig. 1).

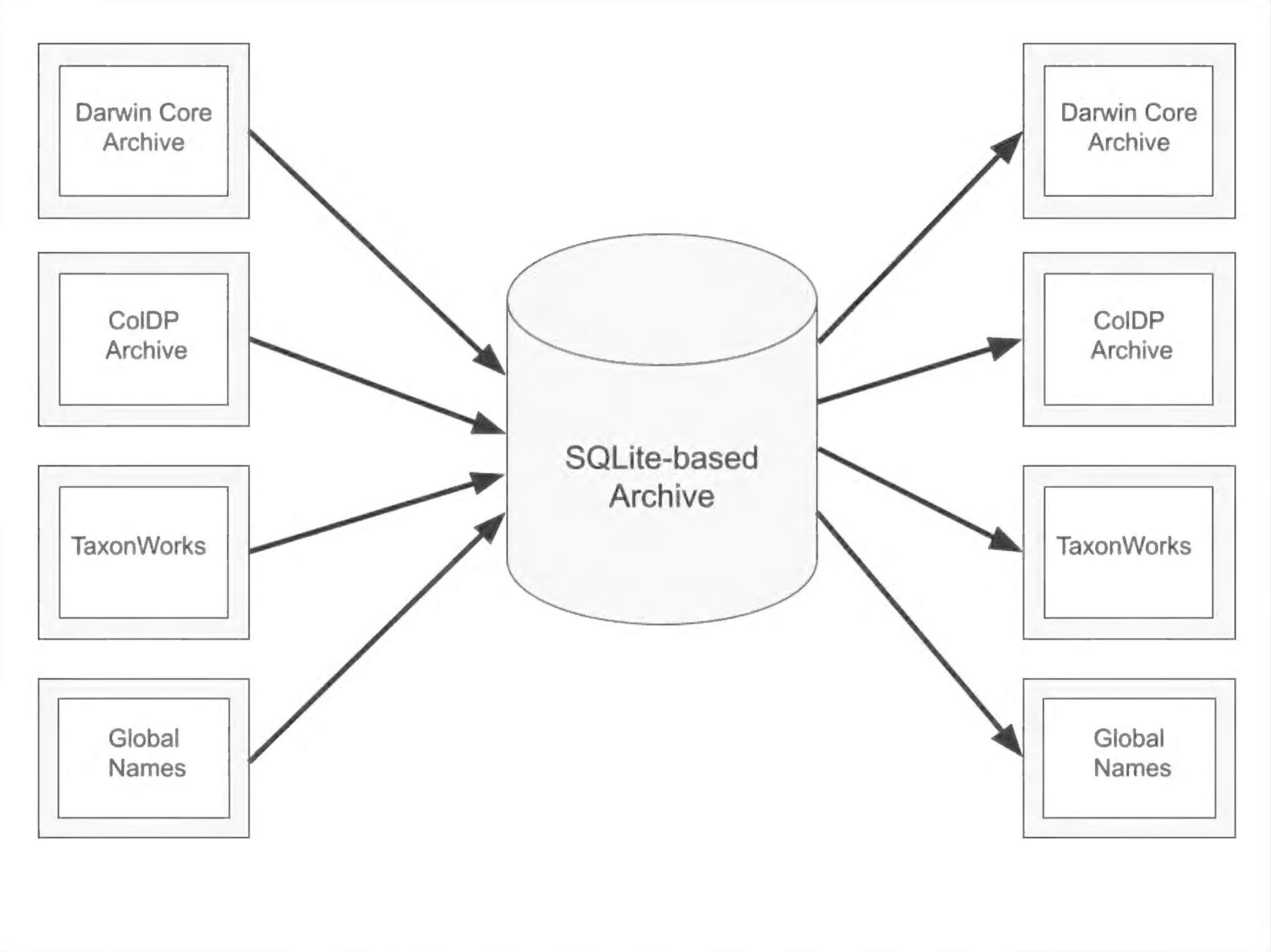

Figure 1.

SQLite-based archive as an universal converter.

2. Semantic versioning of the archive

Like any other data format, the intermediate data exchange SQLite schema will evolve over time. Using Git tags and branches, it is straightforward to follow a semantic versioning paradigm (Semantic Versioning 2024). For example, in "v1.2.5" the numbers respectively indicate major, minor, and patch versions. Major versions include substantial changes that might not be backwards compatible, while minor versions are backwards compatible and patch versions indicate tiny bug fixes.

3. Data querying

With SQLite-based archives being relational databases, they can immediately be queried without any additional parsing and importing. This allows the archives to be explored with ease by anyone familiar with SQL. The Carpentries also provide online learning resources and courses for researchers to learn SQLite (Martinez and Poisot 2017).

4. Biodiversity application development

As SQLite-based archives are relational databases, and the intermediate data exchange schema is standardized and stable, a database file can be used directly by a variety of applications. For example, it would be possible to develop a web application frontend that utilizes the SQLite archive as a backend. It could be possible to develop an ecosystem of biodiversity informatics applications based on SQLite archives.

We believe that adopting SQLite for biodiversity data exchange can reduce operational friction, enhance data accessibility, and promise a streamlined and universally compatible data management framework.

Keywords

data management, standards, data carpentry

Presenting author

Dmitry Mozzherin

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Martinez PA, Poisot T (Eds) (2017) Data Carpentry: SQL for Ecology lesson. http://www.datacarpentry.org/sql-ecology-lesson/. Accessed on: 2024-7-12.
- Semantic Versioning (2024) https://semver.org/. Accessed on: 2024-10-01.
- SQLite (2024) https://www.sqlite.org/. Accessed on: 2024-10-01.
- United States Library of Congress (2024) Recommended Formats Statement. https://www.loc.gov/preservation/resources/rfs/RFS%202024-2025.pdf. Accessed on: 2024-7-12.